最善主義と完璧主義

北本 朝展（KITAMOTO Asanobu）

ROIS-DS人文学オープンデータ共同利用センター（CODH）

国立情報学研究所

http://codh.rois.ac.jp/

ROIS-DS CODHの研究領域

http://codh.rois.ac.jp/

データ駆動型人文学

- 情報学や統計学の考え方を人文学に導入し、人文学の研究方法を革新すること。
- 技術の新しさよりも、人文学における「デジタル変革」を起こすことが重要。

人文学ビッグデータ

- 人文学分野で生み出される大規模データを非人文学に導入し、革新的な研究課題や知識を得ること。
- 研究テーマの「価値」にも、人文学の知見を取り込む。

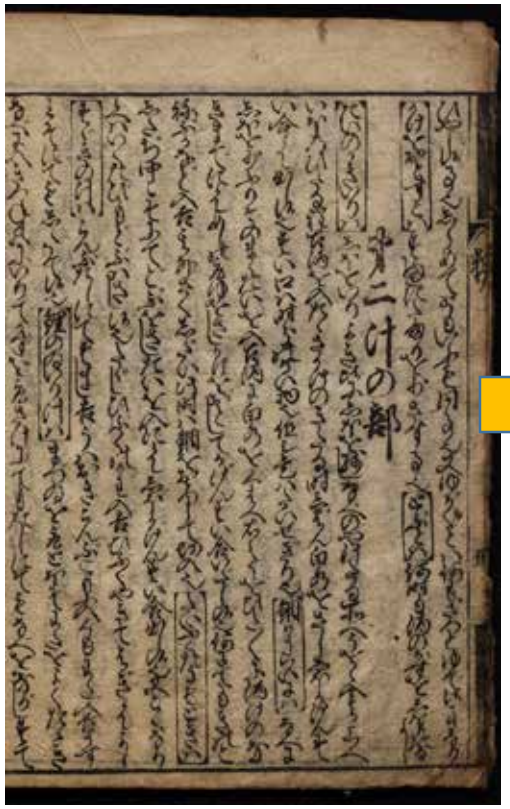
KuroNet

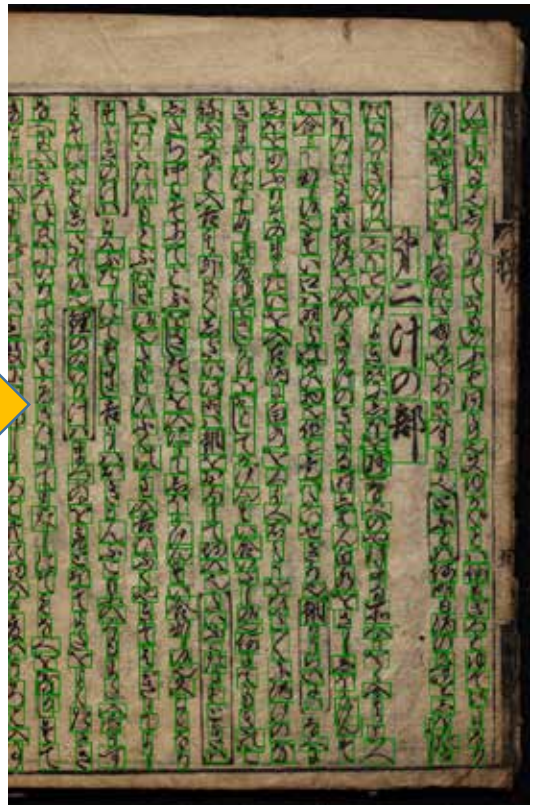
- KuroNetは深層学習を用いたくずし字認識モデル。
- 画像中のすべてのくずし字を現代日本語文字に変換。
- 国文研作成・CODH公開の大規模なくずし字データセット（約100万文字）を学習。
- **カラーヌワット・タリン**（CODH）、**アレックス・ラム**（MIRA）を中心に、国際的な共同研究で開発。

早稲田大学図書館蔵：仮名垣 魯文作『安政風聞集 巻之上』安政3年（1856）

伝統的なOCR手法

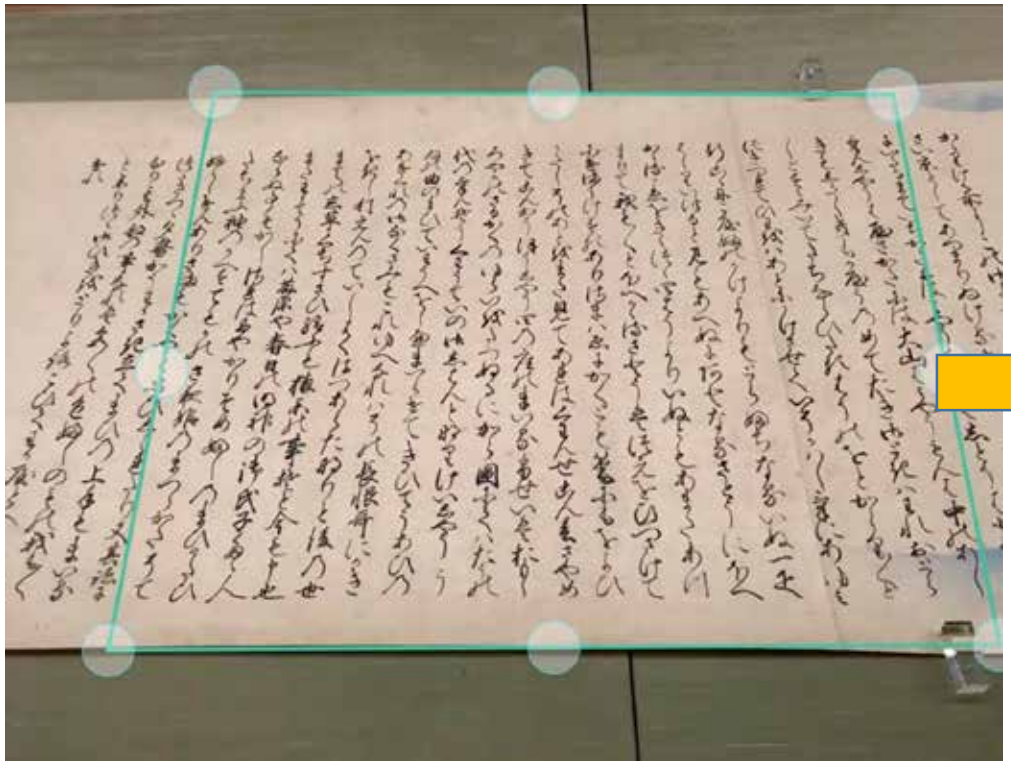
画像 → レイアウト解析 → 文字分割 → 文字認識

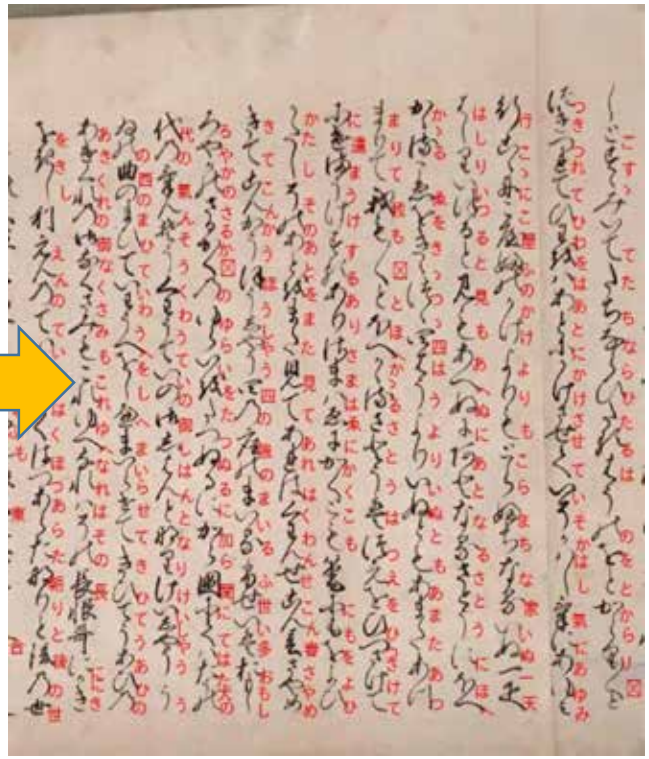




新しいOCR手法（物体検出）

画像 → 文字認識





IIIFを用いたKuroNetサービス

http://codh.rois.ac.jp/kuronet/

kaggle くずし字認識コンペ

機械学習エンジニアが300万人以上登録する、世界最大のデータサイエンスプラットフォームKaggleにて、くずし字認識コンペを開催。

- 期間：7月19日～10月14日まで（約3ヶ月）
- 参加チーム数：293
- 参加者数：338
- 結果提出回数：2652

kaggleコンペの結果

上位は約95%

1. さまざまな国の機械学習エンジニア、研究者が上位入賞。
2. 日本語やくずし字の知識がなくても開発可能。
3. コンペの準備を滞りなく進めるには、情報学者と人文学者の協働が不可欠。

#	Δpub	Team Name	Notebook	Team Members	Score	Entries
1	—	tascj			0.950	13
2	—	Konstantin Lopuhin			0.950	60
3	—	Kenji			0.944	161
4	▲1	YoudaoOCR			0.942	49
5	▼1	See--			0.940	42
6	—	abc			0.939	15
7	—	K_mat			0.934	20
8	—	t-hanya			0.920	21
9	—	Ollie, Nanashi, and Tom			0.910	35
10	—	Zenkei_R&D			0.903	144
11	—	masayai			0.903	12
12	▲5	Kirill Brodt (shad nsk)			0.901	4
13	▲1	James Day			0.901	33
14	▼1	NEU			0.900	54
15	▼3	s tatsuya			0.900	29

コンペが意味すること

- **くずし字の知識なく上位入賞できたのは、そのようにコンペを設計したから！**
- コンペの準備は、情報学者だけでは不可能で、人文学者の専門知識が必須。
- コンペ参加者は、主催者から与えられた評価基準を信じ、ひたすら最適化する。
- **評価基準（価値観）が不適切だと、研究が間違った方向に進んでしまう。**

文字認識から史料批判へ

1. 文字認識：何の文字が書かれているかを、字形だけから推定する。
2. 翻刻：言語としての意味や読みやすさを考慮し、最適な文字を決定する。
3. 史料批判：多くの知識を参照しながら、テキストの内容を適切に理解する。
4. 現代語訳：古文・漢文を現代の日本語に翻訳する。

華北交通アーカイブプロジェクト

北本朝展†1†2　西村陽子†3　貴志俊彦†4　石川禎浩†5

†1 ROIS-DS人文学オープンデータ共同利用センター（CODH）

†2 国立情報学研究所

†3 東洋大学

†4 京都大学東南アジア地域研究研究所

†5 京都大学人文科学研究所

2019年2月公開

華北交通アーカイブ

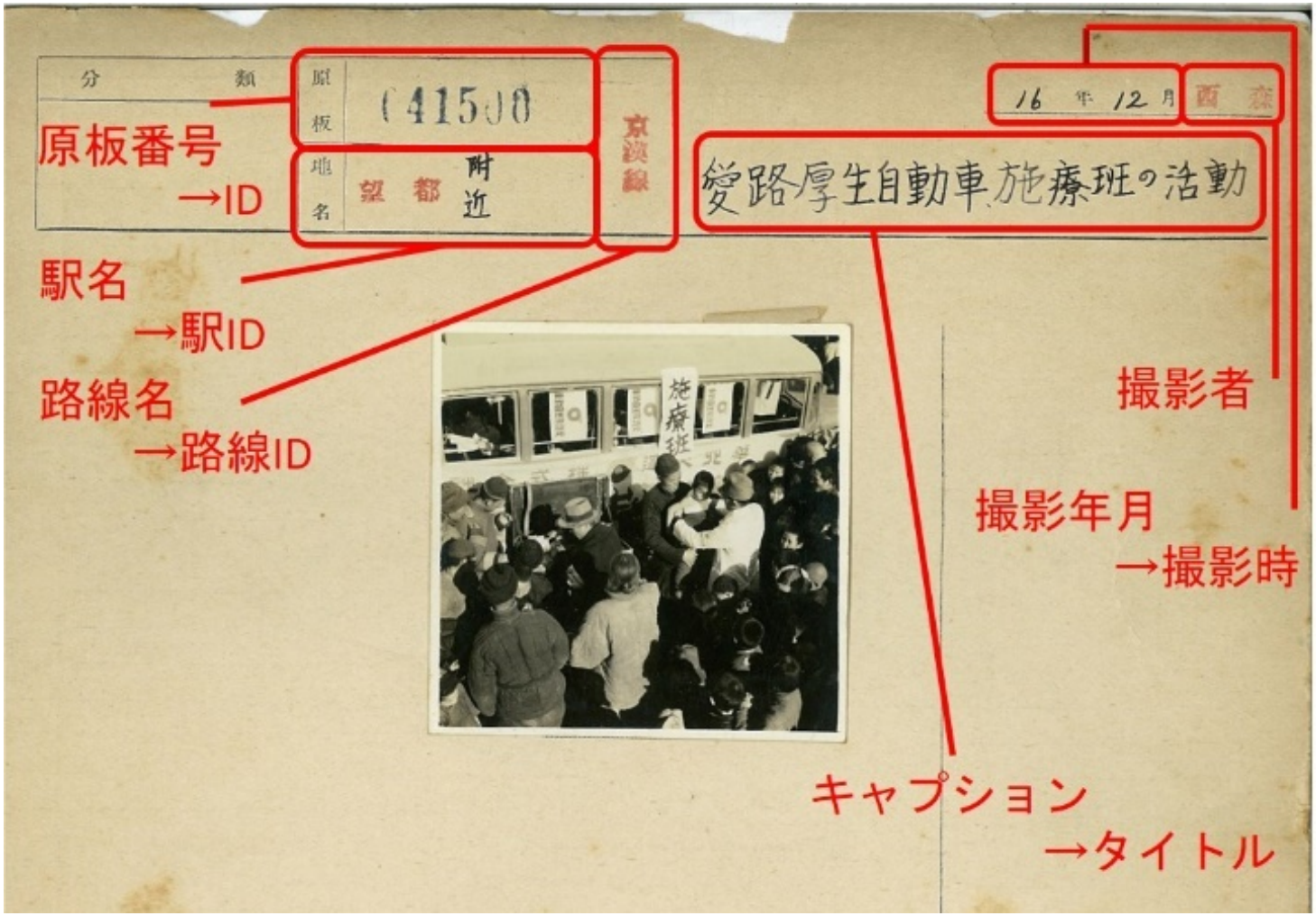
http://codh.rois.ac.jp/north-china-railway/

よみがえる膨大な白黒写真 - 国策鉄道会社が遺した戦時期広報用写真の研究データベース

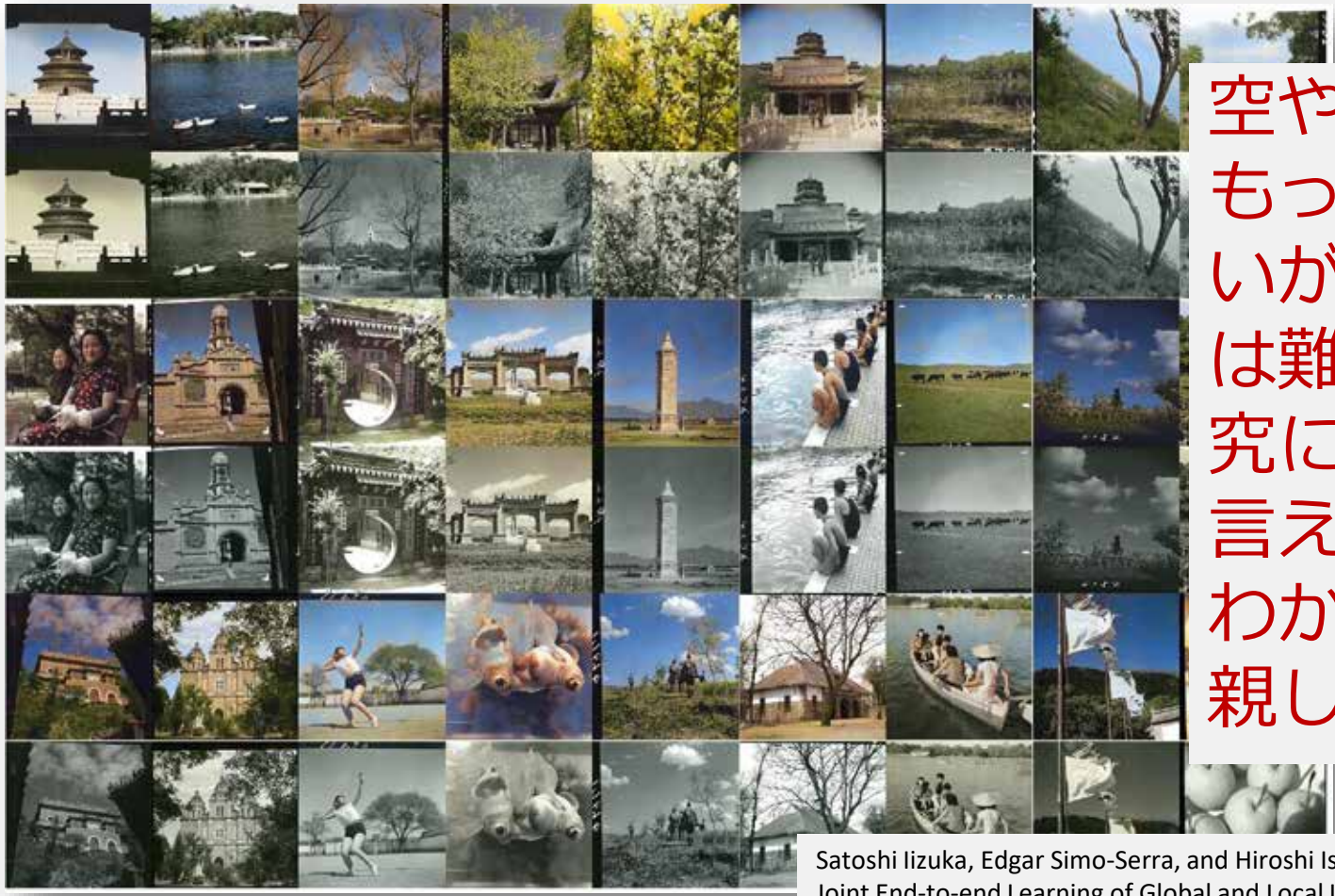
公開目的・公開宗旨: 日本語 | 簡体中文 | 繁体中文 |

華北交通アーカイブは、日中戦争期に中国北部・西北部(華北)の交通インフラを管轄していた華北交通株式会社(以下、華北交通)が所蔵していた3万8千点あまりの広報用ストックフォトをもとに構築された統合型の研究データベースです。ここに含まれる個々の写真は、当時の風物を詳細に調べるための精密(precise)な資料としての価値があります。一方、写真群全体の意味を考えるには、特定の撮影対象がなぜ選ばれたのかという意図を考えることも不可欠であり、そうしたバイアスを踏まえると、この写真群は事実の正確(accurate)な描写とはみなせない面もあります。こうした問題を考えるために、戦時期日本人／日本語話者向けに用いられた広報用写真を学術研究資源として公開しました。今後もその全貌について、日中間はもとより、国際的な枠組みのもとで調査していきます。[もっと詳しく..] [華北交通 路線図..]

華北交通写真の台紙（旧版）



画像の自動カラー化



空や植物はもっともらしいが、人工物は難しい。研究に有効とは言えないが、わかりやすく親しみやすい。

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. "Let there be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification". ACM Transaction on Graphics (Proc. of SIGGRAPH), 35(4):110, 2016.

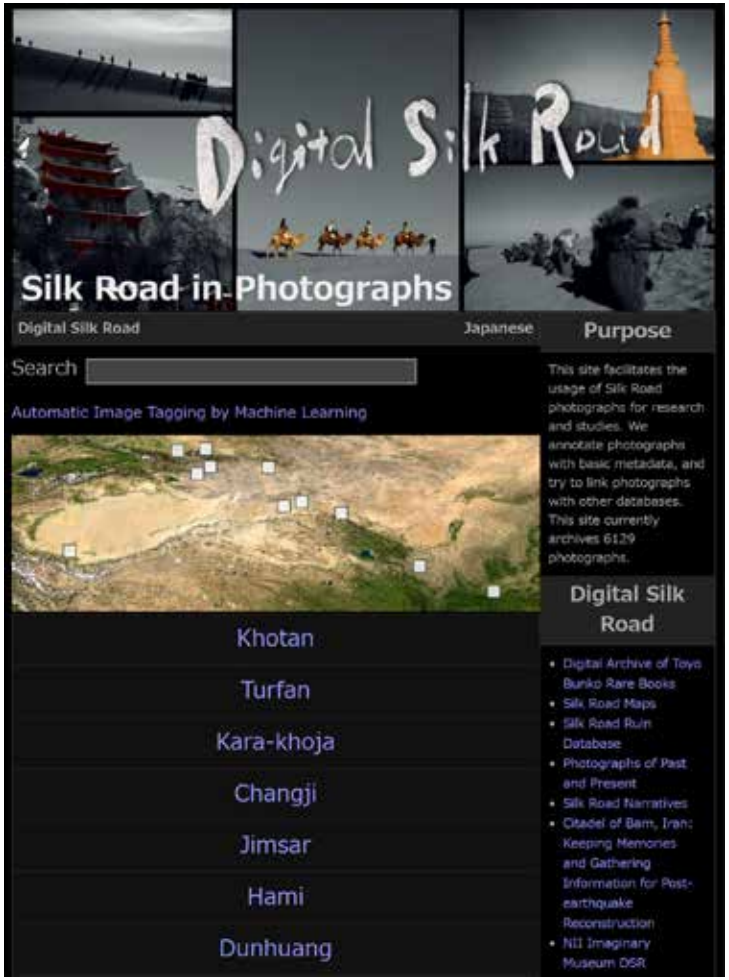
画像の自動タグづけ

メタデータがない画像の検索補助として、機械による自動タグ付けを利用。必ずしも正確ではないが、うまく使えば意外な画像を探せる。

ILSVRC2012データセットを対象に1000クラスの画像分類を学習したTensorFlow実装のInception V3モデルを利用。

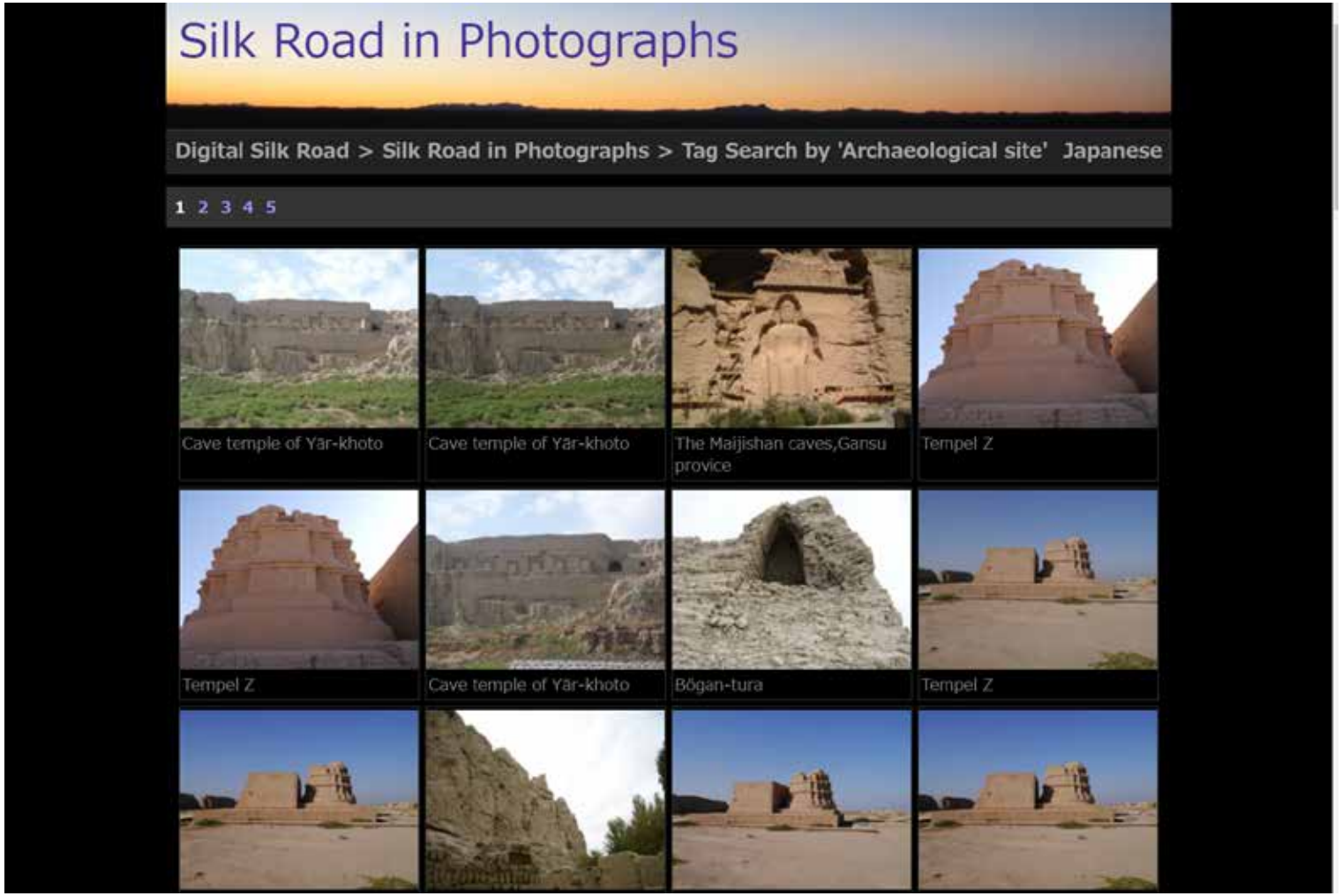
フィールドワーク現地写真

http://dsr.nii.ac.jp/photograph/



1. シルクロード考古遺跡のフィールドワークによる現地撮影写真コレクション。
2. 長期間にわたる6,129枚の写真。
3. 東洋大学の西村陽子氏を中心とした協働により構築。

タグ：Archaeological Site



タグ： Snow

Silk Road in Photographs
Digital Silk Road > Silk Road in Photographs > Tag Search by 'Snow'
Japanese
1 2 3 4
Ruined town of Nanhu（Ancient city of Shou'chang）
Ruined town of Nanhu（Ancient city of Shou'chang）
Ruined town of Nanhu（Ancient city of Shou'chang）
Ruined town of Nanhu（Ancient city of Shou'chang）
Ruined town of Nanhu（Ancient city of Shou'chang）
Ruined town of Nanhu（Ancient city of Shou'chang）
Ruined town of Nanhu（Ancient city of Shou'chang）
Ruined town of Nanhu（Ancient city of Shou'chang）

タグ：Wood

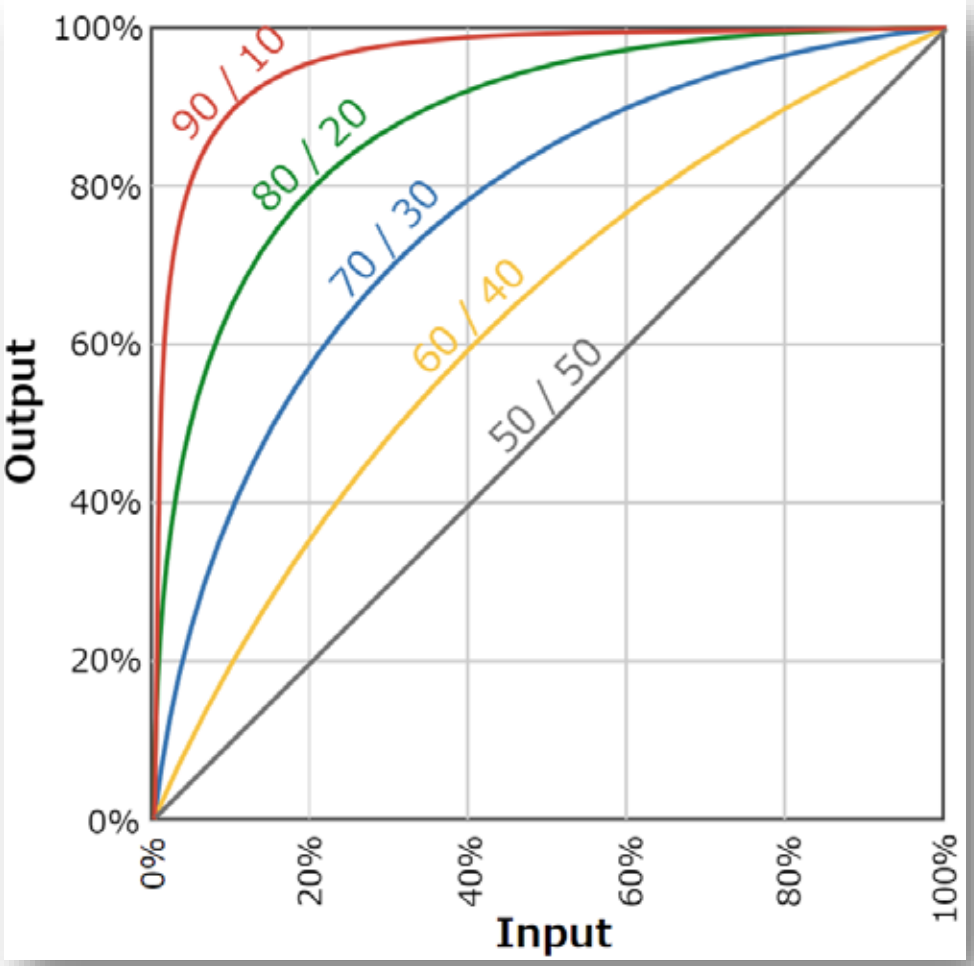
Silk Road in Photographs
Digital Silk Road > Silk Road in Photographs > Tag Search by 'Wood'
Japanese
1 2 3 4
Khohmari caves
Ruined buddhist shirines, Ara-tam
Tempel Z
Tempel Z
Tempel Z
Tempel K
Tempel Z
Tamguluk-tāsh

一般名詞と固有名詞

1. 自然物を対象とした一般名詞的なタグ付けであれば、ある程度は利用可能。
2. 人名、地名、商品名など固有名詞的なタグ付けには、専用のAIが必要。
3. 専門的な研究のための詳細な分類には、より多くのデータが必要になる。
4. 背景知識を適用し文脈を読むこと、意味的な重みづけを行うことは難しい。

80:20法則と100%への道

1. AIは80:20法則の考え方において有効なツール。
2. 最終的な目標が100%だと、その後がしんどい。
3. AIを使わない方が効率的となる場合もある。



AIは人間の代替か？

1. 最善主義（best effort）：人間の苦痛を軽減するツールとしての価値は高い。
2. 完璧主義（perfection）：例外を扱って人間並みに精度を高めるのは困難。
3. 「苦痛の軽減」と「価値の増大」：前者の一部はAIで代替される可能性あり。
4. 研究レベルの向上：AIが苦痛の軽減に役立つなら、研究者は本来やるべきことに集中し、価値の増大に貢献できる。